COMPLIANCE COMPONENT

DEFINITION	
Name	Extract Transform & Load (ETL) Best Practices
Description	In defining the best practices for an ETL System, this document will present the requirements that should be addressed in order to develop and maintain an ETL System. These best practices will address the constraints placed on the ETL system and how best to adapt the ETL system to fulfill the requirements. NOTE: The functionality of relational databases is rapidly eliminating the ETL category by incorporating transformation functionalities. This is creating a new process ELT (extract load and transform) where all complex processing of data occurs inside the database itself.
Rationale	The requirements and constraints placed on an ETL system must be addressed up front and not during the implementation of the ETL process. When these best practices are addressed and decisions made in the planning process, there is a better chance that the ETL implementation process will be successful.
Benefits	Presenting the best practices for meeting the requirements of an ETL system will provide a framework in which to start planning and/or developing the ETL system which will meet the needs of the data warehouse and the end-users who will be using the data warehouse.

ASSOCIATED ARCHITECTURE LEVELS	
Specify the Domain Name	Information
Specify the Discipline Name	Knowledge Management
Specify the Technology Area Name	Extract Transform and Load (ETL)
Specify the Product Component Name	

COMPLIANCE COMPONENT TYPE	
Document the Compliance Component Type	Guideline
Component Sub-type	

COMPLIANCE DETAIL	
State the Guideline, Standard or Legislation	**Business Needs** For the purposes of an ETL system, business needs can be narrowly defined to mean the information content that end users need to make informed decisions. This information content will be extracted from the appropriate available sources and output to a data warehouse or data store for a business application which is accessible to the end-user. **Compliance Requirements** The ETL system must plan for the compliance requirements placed on it by outside sources. State and Federal rules, regulations, standards and guidelines must be taken in consideration during the planning stage. Although the rules, regulations, standards and guidelines will most likely apply to the data warehouse or business application to which the ETL will transfer data, the ETL process itself must account for the compliance requirements so that the data transferred is as usable as possible.

Metadata

Metadata has many different meanings depending on its context. Several metadata elements associated with the ETL process are valuable to the data warehouse or user application and must be provided to the end users.
Types of metadata derived during the ETL process include the following:

Data Lineage – also known as logical data mapping, illustrates transformations applied to a data element between its original data source and its ultimate data output.

Business Definitions – Every table created in the ETL process stems from a business definition. Although these definitions can be captured during the ETL process, these definitions will be carried forward to the data warehouse or application for which the end user intends to use them.

Technical Definitions – Technical definitions describe the physical attributes of data elements, including the structure, format, and location. Properly document technical metadata for all ETL tables to minimize ambiguity and ensure usability.

Process Information – Processes that load ETL tables must record their statistics along with statistics of the data warehouse table loads. Although information about ETL data loads need not be presented to the end user, the ETL administrator must know exactly how many records were loaded into each table, with success and failure statistics for each process. A measure of data freshness is useful both for ETL administrators and end users.

ETL tools and data-modeling tools offer metadata capture capabilities that would otherwise cause metadata documentation to be a tedious, labor intensive task. It is estimated that the structure-oriented metadata captured by these tools is around 25% of the total metadata captured in the ETL process. Another 25% of the metadata captured describes the results of data cleansing. The remaining 50% of the metadata consists of process information. It is essential that the ETL tool provides this type of metadata.

Data Profiling

This is a systematic review of the quality, scope and context of the data source from which the ETL system will extract data. A 'clean' data source will require minimal transformation, while a 'dirty' data source will require extensive data transformation. Data profiling may ultimately reveal that the 'dirty' data source cannot be supported by an ETL process. It is better to discover such issues in the planning stage rather than during implementation of the ETL system.

Security Requirements

Security for an ETL System can be simplified by establishing an overall rule that end-users cannot be allowed to access any data during the ETL process. Access to the ETL system should be limited to only those who need access to make the ETL process work. Security for the data which is output to the data warehouse, which the ETL process supports, should be addressed as a data warehouse issue and not as part of the ETL process.

Data Integration

For ETL, data integration takes the form of conforming dimensions and conforming facts. Conforming dimensions establish common dimensional attributes across separate databases so the drill-across reports can be generated

	using these attributes. Conforming facts are the common business metrics, such as key performance indicators across separate databases, which allow numbers to be compared mathematically by calculating differences and ratios. NOTE: Newer approaches to these functions have been labeled ELT (extract, load, transform). With this approach the data is transformed on the target after being loaded. This is especially pertinent if the target database is powerful enough where it can be used to perform all transformations and optimize both performance and investment. In this case there is no 'useless' data transferred on the network and takes full advantage of the power of the RDBMS. As is the case with all such tools the flexibility to switch to the more traditional ETL architecture is always there. **Data Latency** Data latency refers to how quickly the data must be delivered to the end users. Data latency will have a significant impact on the design and implementation of the ETL system. Traditional batch-oriented data flows may be sufficient to meet the needs of the business. However, if time is critical, the ETL process may have to be implemented using a streaming data flow. **Archiving and Lineage** ETL data should be stored at each step where a major transformation of data has occurred. In a basic data flow, this would occur after each step: extract, clean, conform, and deliver. Once the data is stored to a media, it must be determined if the stored data is to be archived and if so, for what length of time. If there are legal requirements, these must be met. If there are not, then it should be determined in the planning stage of the system. In most cases, it is easier to restore data from a permanent media than to reprocess the data thru the ETL system. **End User Delivery Interfaces** The final step in the ETL process is the transfer of data to a data store. This may be an application, data warehouse or data mart. Regardless of the final destination of the data, it should be the responsibility of the ETL system to transfer the data to the data store in the most usable format possible. This should simplify the application development process and leave the process of data transformation to the ETL system rather than a business application. **Legacy Systems** An ETL system should not be developed with 'legacy' software. Even if this 'legacy' software is enhanced to provide ETL capabilities, it should not be our practice to propagate the development of large scale processes using software which is not considered to be 'current' software.
Document Source Reference #	*The Data Warehouse ETL Toolkit* by Ralph Kimball

Compliance Sources			
Name		*Website*	
Contact Information			
Name		*Website*	
Contact Information			

KEYWORDS	
List Keywords	Extract Transform and Load (ETL), Extract Load and Transform (ELT), Extract Transform and Transport (ETT), Extract Transform and Move (ETM), Data Warehouse, Data Mart, Data Cleansing, Extraction, Enterprise Application Integration, Enterprise Service Bus, Data Exchange, Information Exchange, Metadata

COMPONENT CLASSIFICATION				
Provide the Classification	☐ *Emerging*	☒ *Current*	☐ *Twilight*	☐ *Sunset*
Sunset Date				

COMPONENT SUB-CLASSIFICATION		
Sub-Classification	Date	Additional Sub-Classification Information
☒ *Technology Watch*	6-7-06	Traditional ETL approaches rely on proprietary ETL engines deployed between sources and targets. The functionality of relational databases is rapidly eliminating the ETL category by incorporating transformation functionalities. This is creating a new process ELT (extract load and transform) where all complex processing of data occurs inside the database itself. See ELT Reference document.
☐ *Variance*		
☐ *Conditional Use*		

Rationale for Component Classification	
Document the Rationale for Component Classification	

Migration Strategy	
Document the Migration Strategy	

Impact Position Statement	
Document the Position Statement on Impact	

CURRENT STATUS				
Provide the Current Status	☐ *In Development*	☐ *Under Review*	☒ *Approved*	☐ *Rejected*

AUDIT TRAIL			
Creation Date	04/05/06	*Date Approved / Rejected*	06/13/06
Reason for Rejection			
Last Date Reviewed		*Last Date Updated*	
Reason for Update			